DOCUMENT RESUME

ED 162 663 IR 006 668

AUTHOR Pepinsky, Harold B.
TITLE A Computer-Assisted Language Analysis System (CALAS) and Its Applications.
PUB DATE 78
NOTE 16p.; For related document, see ED 090 948

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Case (Grammar); Componential Analysis; *Computational Linguistics; Computer Programs; Data Analysis; Data Processing; *Discourse Analysis; Error Analysis (Language); Language Patterns; *Linguistic Patterns; Sentence Diagraming; *Sentence Structure; Speech Communication
IDENTIFIERS *Computer Assisted Language Analysis System

ABSTRACT

A Computer-Assisted Language Analysis System (CALAS) was developed as a syntactic and semantic analyzer of machine readable text in English. CALAS includes a set of computer programs, an algorithm for implementation, and human editors who assist the computer and its programmer in the processing of data. Data analysis is accomplished in three stages: (1) syntactic analysis of text, identifying each work in sequence in terms of its grammatical equivalent; (2) aggregation of the individual words into phrases identified in terms of their grammatical equivalents; and (3) aggregation of phrases into clauses, with component phrases identified in terms of the roles each plays and exhibited to display a main or independent clause. Discussion of the literature focuses on the relative frequencies with which the different types of verb phrases are used, and the measure of structural or stylistic complexity. (JEG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

- ED162663

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A COMPUTER-ASSISTED LANGUAGE ANALYSIS SYSTEM (CALAS) AND ITS APPLICATION

by

Harold B. Pepinsky

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Harold B. Pepinsky

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

IR004668

I R

A COMPUTER-ASSISTED LANGUAGE ANALYSIS SYSTEM (CALAS) AND ITS APPLICATIONS

Harold B. Pepinsky

The Ohio State University

A Computer-Assisted Language Analysis System, or CALAS, was developed to help its users understand and interpret human talk. As my contribution to this symposium, I shall describe briefly what eight years of research on and by means of CALAS are beginning to tell us. First, CALAS itself will be described. Second, I shall outline a pivotal research study, just submitted for publication, which provides construct validation for CALAS. Then, I shall refer to other work, illustrative of CALAS's applications to psychological and educational research.

First, CALAS. CALAS ingests, as its raw material, what is called in the trade, machine-readable text in the English language. What that means is the original text or transcriptions of connected speech or writing in English, key-punched onto computer cards or tapes. As such, CALAS is in the vanguard of a current flurry of interest in what is popularly known as "discourse analysis" (e.g., Freedle, 1977). CALAS is both a syntactic and a semantic analyzer of the texts it ingests. Initially, that is, CALAS makes a linear translation of the discourse into its grammatical elements. It does this by reference to the sequential order in which words appear in the lines of a text. Then, without necessary regard for word order, CALAS identifies relationships existing within meaningful clusters of words. Let me specify a bit more how all this is accomplished.

By design, CALAS includes a set of computer programs and an algorithm--or set of rules for implementing the programs, on the one hand, and human editors who, according to instructions, assist the computer and its human

8/22/78

programmer in the processing of data. With these resources, data analysis is accomplished in three stages. It is in Stage 1, called EYEBALL, that CALAS makes the syntactic analysis of text, identifying each word in sequence in terms of its grammatical equivalent, e.g., noun, verb, adjective, adverb, preposition. CALAS does this by recourse to a small dictionary (of approximately 600 words) and, as importantly, a set of rules for identifying other words in terms of where in sequence each appears and what role it is thus intended to play. Where alternative roles are plausible for a word, the computer is programmed to list these in the order of their most likely occurrence for that word in that slot, e.g., as adjective, verb, noun. At this point, one or more humans (we recommend at least two persons) edit the text according to instructions, rapidly correcting the relatively few evident "errors" made by the computer. That editing is an important addition to the process of analysis because, as will become apparent, earlier errors become compounded in later stages of analysis.

The edited output of EYEBALL becomes input for a second stage of analysis called PHRASER. Here, guided by a second set of programs, the computer aggregates the individual words into phrases, which are again identified in terms of their grammatical equivalents, e.g., as noun phrases, verb phrases, adverbial phrases, prepositional phrases; also, as conjunctions and subordinators (i.e., terms that introduce main and subordinate or partial clauses). PHRASER, like EYEBALL, is then edited, and the system is ready for its third and--at present--final stage of analytic display.

Stage 3 in the analysis of data processed by CALAS is called CLAUSE/CASE. At this stage, the computer is instructed to do three things, and by reference to a third set of programs. First, phrases are aggregated into clauses, with the component phrases displayed within each clause. Second,

the phrases within a clause are identified in terms of the roles each plays within the clause. Because by definition a clause contains one and only one predicate--notably, a verb phrase--the verb phrase itself becomes an essential feature of the clause, with other phrases as optional features. Verb phrases, then, are identified as particular types, basically as verbs of state, action, process, or action-process, and secondarily as compounds of experiential or benefactive states or actions. Noun phrases that accompany the verb phrase are thus identified in terms of their case roles within a clause, as objects, agents, experiencers, or beneficiaries of a state or activity. Finally, the clauses themselves are exhibited to display a main or independent clause, along with clauses subordinate to it and in the order of their embedding within the block of clauses. At the present time, the computer's unedited display of clauses and phrases within clauses appears to be highly accurate; a greater amount of human assistance is needed in identifying phrase roles within a clause and the relative embeddedness of clauses within blocks.

My account of how CALAS is used to identify and display information for analysis has had to be abbreviated. Details on the development of the system, its rationale, a research strategy based upon its use, and a detailed manual of procedures for persons who want to use it may be found elsewhere (Cook, 1972a, 1972b, 1975; Meara, 1976; Patton, Fuhriman, & Bieber, 1977; Pepinsky, 1974, in preparation; Rush, Pepinsky, Meara, Landry, Strong, Valley, & Young, 1974; Strong, 1974; Young, 1973). A number of reports, describing research applications of CALAS and how information yielded by CALAS has been further analyzed to provide quantitative assessments of various kinds of discourse, is also available (e.g., Baker, 1976; Baker, Sandman, & Pepinsky, 1975; Bieber, Patton & Fuhriman, 1977; Bieber, 1978; Hurndon, Pepinsky, & Meara, 1978; May, 1977; Murray, 1978; Patton, Fuhriman, & Bieber, 1977; Shannon, 1977).

To date, two major kinds of information have been quantified. The first includes content measures: saliently, the relative frequencies with which the different types of verb phrases are used. The second includes measures of structural or stylistic complexity; for example, the ratio of number of clauses to main clauses as a measure of average block length, measures of the average embeddedness of clauses within blocks, and the ratio of the number of verb phrases--as essential relaters within clauses--to the number of so-called peripheral relaters--adverbial and prepositional phrases that modify verbs.

Using measures of stylistic or structural complexity, Hurndon et al. (1978) have completed a study that provides an important construct validation for CALAS. The research centered on a linguistic analysis of 39 protocols used in scoring the Paragraph Completion Method (PCM) and to obtain a measure of what its principal author, David Hunt, identifies as "conceptual level." The construct has to do with how people are inferred to structure their experiences. Persons scoring high on the Paragraph Completion Method, as a measure of Conceptual Level, are labeled as "cognitively complex"; those with low scores, as "cognitively simple" (cf. Hunt, Butler, Noy, & Rosser, 1976). The 39 protocols consisted of exemplary paragraphs written by research subjects to complete six stems with which to initiate sentences; each set of responses illustrates a different score based on raters' assessments of an illustrative subject's conceptual level. These marker ratings of each protocol for conceptual level could now be compared with linguistic measures of structural complexity obtained by means of CALAS. Individual correlations and stepwise multiple regression analyses indicated that simple frequency measures of the numbers of words, phrases, and clauses used in written response to the PCM accounted for much of the variance in Conceptual Level scores. Results of a factor analysis supported this finding, but suggested that something like an index of moral judgment also might be involved. A

number of earlier validation studies (cited in Hunt et al., 1976) indicate the utility of the CL measure in teaching and counseling situations. Measures of stylistic complexity based on the use of CALAS--especially, the sheer amount of writing involved--may account more parsimoniously for these findings.

As indicated earlier, CALAS has already been applied to a variety of problems in everyday life. For example, several investigations by Michael Patton and his colleagues (e.g., Bieber, 1978; Bieber, et al., 1977; Patton, et al., 1977) give evidence of convergence and tracking behavior on the part of clients in counseling that takes place over repeated occasion. Here, "convergence" implies that, over time, therapist and client become more alike in their language. "Tracking" implies that as the one shifts in usage--e.g., from talking about feelings to talking about states--so does the other. Gregory May's (1977) collaborative research supports this general finding. A comparable kind of convergent behavior is revealed by the client, Gloria, in her celebrated brief interviews with Carl Rogers, Fritz Perls, and Albert Ellis. Even within her brief interviews, Gloria is differentially responsive to her three therapists, as determined by changes in measured content and style (Murray, 1978; Shannon, 1977). For example, in conversation with Rogers, she comes to talk more about states--the way things are. In conversation with Ellis, she talks more about taking action.

Let me make this clearer by reference to a further analysis of the Gloria data, which Murray (1978) has just completed. In Figure 1 of your hand-out, the percentages of verb usage by Carl Rogers, Albert Ellis, and Fritz Perls, along with the verb usage by Gloria, are displayed. Figure 2 identifies and defines the five types of verb shown on the axes of the Charts in Figure 1: S=Stative, SEC=Stative Experiencer/Cognitive, SEA=

Stative Experiencer/Affective, A=Agentive (a composite of Action and Action/Process), and AE=Agentive Experiencer. Note that Rogers, over two occasions, uses mainly State verbs, i.e., talks about the way things are and the shape of Gloria's frequency curve tends slightly to be like his. Perls, in contrast, uses many more verbs of agency and of verbs linking an agent with an experiencer. Again, there is some tendency for Gloria's verb usage to resemble that of Perls. Finally and dramatically, however, it is Ellis who talks about making something happen, exemplified in his most frequent employment of verbs of agency; secondarily, he talks about states. Gloria's profile of verb usage clearly parallels that of Ellis. Figure 3 shows that over two occasions, as indicated by excerpts from talk early in the interviews and toward the middle, the protagonists remain consistent in the relative frequencies with which they employ the verb types in their counseling. It is notable that Rogers, like two other client-centered therapists whose talk we have analyzed by means of CALAS, talks less about feelings than might have been expected and is more apt to describe the way things are. Shannon's data (not shown, but in Shannon, 1977) indicate Rogers' speaking style to be relatively uncomplicated, while Ellis, the exhorter, uses much more complicated talk.

Figures 1-3 about here

In two laboratory experiments, Morris Baker (1976; also Baker, Sandman, & Pepinsky, 1975) measured changes in the amount and kind of talk exhibited by subjects under different levels of physiological arousal. Though interactions of style and content with physiological arousal were found, a consistent result from experiment to experiment was that sheer amount of talk was positively related to amount of arousal. Sheer volume of utterance, you will recall, accounted also for much of the variance in measured Conceptual Level. The counseling research not only reveals differential amounts of

talk by different therapists, but evidence of tracking on the part of the client--covariation over time with the counselor--and, in some cases, a convergence of the client toward the level of talk manifested by the therapist.

Kathleen McCarthy has underway a study of sex role stereotyping that may be exhibited by television interviewers and persons interviewed and as revealed in like sex and opposite sex interviews analyzed by means of CALAS. Although much has been written about sex role stereotyping, relatively little pertinent or reliable information has been obtained in the analysis of conversations between or within the sexes.

Much remains to be done. My colleagues and I believe, however, that substantial progress has been made, by means of CALAS, toward interpreting human discourse. CALAS provides both a sensible method of analyzing human talk and, as indicated to date, a valuable source of information about it.

REFERENCES

Baker, W. M. Physiological and linguistic responses to varied states of arousal. Unpublished doctoral dissertation, Ohio State University, 1976.

Baker, W. M., Sandman, C. A., & Pepinsky, H. B. Affectivity of Task Rehearsal Time and Physiological Response. Journal of Abnormal Psychology, 1975, 84, 539-544.

Bieber, M. R. A language analysis of three counseling series. Unpublished doctoral dissertation, University of Utah, 1978.

Bieber, M. R., Patton, M. J., & Fuhriman, A. J. A metalanguage analysis of counselor and client verb usage in counseling. Journal of Counseling Psychology, 1977, 24, 264-271.

Cook, W. A. A set of postulates for case grammar analysis. In Languages and Linguistics: Working Paper, No. 4. Washington, D. C.: Georgetown University Press, 1972a. Pp. 35-49.

Cook, W. A. A case grammar matrix. In Languages and Linguistics: Working Papers No. 6. Washington, D. C.: Georgetown University Press, 1972b. Pp. 15-47.

Cook, W. A. Stylistics: measuring style complexity. Georgetown University Working Papers on Languages and Linguistics, No. 11. Washington, D. C.: Georgetown University Press, 1975. Pp. 106-120.

Freedle, R. O. (Editor) Discourse Production and Comprehension. New Jersey: Alex Publishing Company, 1977, Vol. 1.

Hunt, D. E., Butler, L. F., Noy, J. E., & Rosser, M. E. Assessment of conceptual level: Paragraph Completion Method (PCM). (Revision 2) Unpublished manuscript, Ontario Institute for Studies in Education, 1976.

Hurndon, C. J., Pepinsky, H. B., & Meara, N. M. Conceptual level and structural complexity in language. Department of Psychology, Ohio State University, 1978. (Submitted for publication)

May, G. D. Psychotherapy and language: Linguistic convergence between therapist and client. Unpublished doctoral dissertation, Ohio State University, 1977.

Meara, N. M. A computer-assisted language analysis system for research on natural language. Symposium presentation at the Interamerican Congress of Psychology, Miami Beach, December, 1976.

Murray, W. J. A computer-assisted comparative analysis of specific language structures in the counselor/client relationship. Unpublished manuscript, Department of Psychology, Ohio State University, 1978.

Patton, M. J., Fuhriman, A. J., & Bieber, M. R. A model and a metalanguage for research on psychological counseling. Journal of Counseling Psychology, 1977, 24, 25-34.

Pepinsky, H. B. A metalanguage for systematic research on human communication via natural language. Journal of the American Society for Information Science, 1974, 25 (1), 59-69.

Pepinsky, H. B. A metalanguage of text. In H. Fisher and R. Diaz-Guerrero (Eds.), Language and logic in personality and society. New York: Academic Press (book in preparation).

Rush, J. E., Pepinsky, H. B., Meara, N. M., Landry, B. C., Strong, S. M., Valley, J. A., & Young, C. E. A computer-assisted language analysis system. Columbus, Ohio: Computer and Information Science Research Center, OSU-CISRC-TR-73-9, Ohio State University, 1974.

Shannon, J. W. III. A computer-assisted analysis of the language complexity of psychotherapists with different theoretical orientations. Unpublished honors thesis, Ohio Dominican College, Columbus, Ohio, 1977.

Strong, S. M. An algorithm for generating structural surrogates of English text. Journal of the American Society for Information Science, 1974, 25, (1), 10-24.

Young, C. E. Development of language analysis procedure with application to automatic indexing. Columbus, Ohio: Computer and Information Science Research Center, OSU-CISRC-TR-73-2, Ohio State University, 1973.

FOOTNOTES

Part of a Symposium, H. B. Pepinsky, Chair, on "Applications of Semantic Grammars to Psychological and Educational Research," Southeastern Psychological Association Meetings, Atlanta, Georgia, 17 March 1978.

As Co-Director of the Group for Research and Development in Language and Social Policy in the Ohio State University's Mershon Center, under whose auspices this paper is written, the author gratefully acknowledges the continued support and encouragement of the Mershon Center.

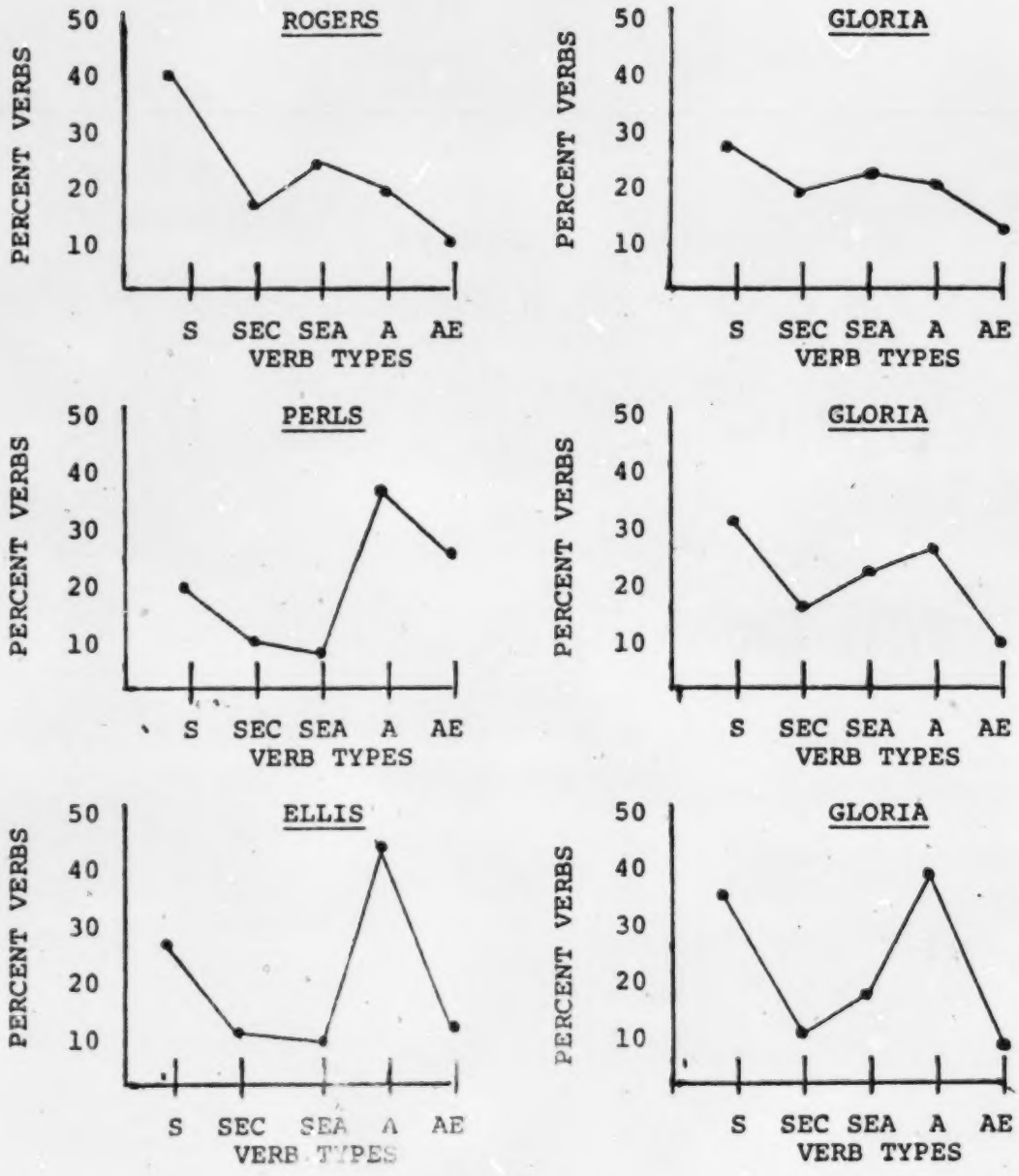

Figure 1. Perce[illegible]erbs used by Rogers, Perls, Ellis, and Gloria. (From Murr[illegible]).

Figure 2. Five Verb Types Used Predominantly by Gloria and Her Therapists in Two Excerpts from Each Interview (see Murray, 1978)

STATIVE (S)--defines a state or property of an object.
e.g., I <u>am</u> sorry.

STATIVE EXPERIENCER/COGNITIVE (SEC)--the verb defines an experiencing person's <u>cognitive</u> state.
e.g., I <u>think</u> I <u>know</u> what you said.

STATIVE EXPERIENCER/AFFECTIVE (SEA)--defines an experiencer's <u>affective</u> state.
e.g., She <u>feels</u> bad.

AGENTIVE (A)--an agent makes something happen, either as ACTION (--does something e.g., The dish <u>ran</u> away with the spoon) or as ACTION/PROCESS (--does something <u>to</u> something, Mary <u>hit</u> John)

AGENTIVE EXPERIENCER (AE)--someone or something acts as an agent to give someone an experience.
e.g., The weather <u>troubled</u> us.

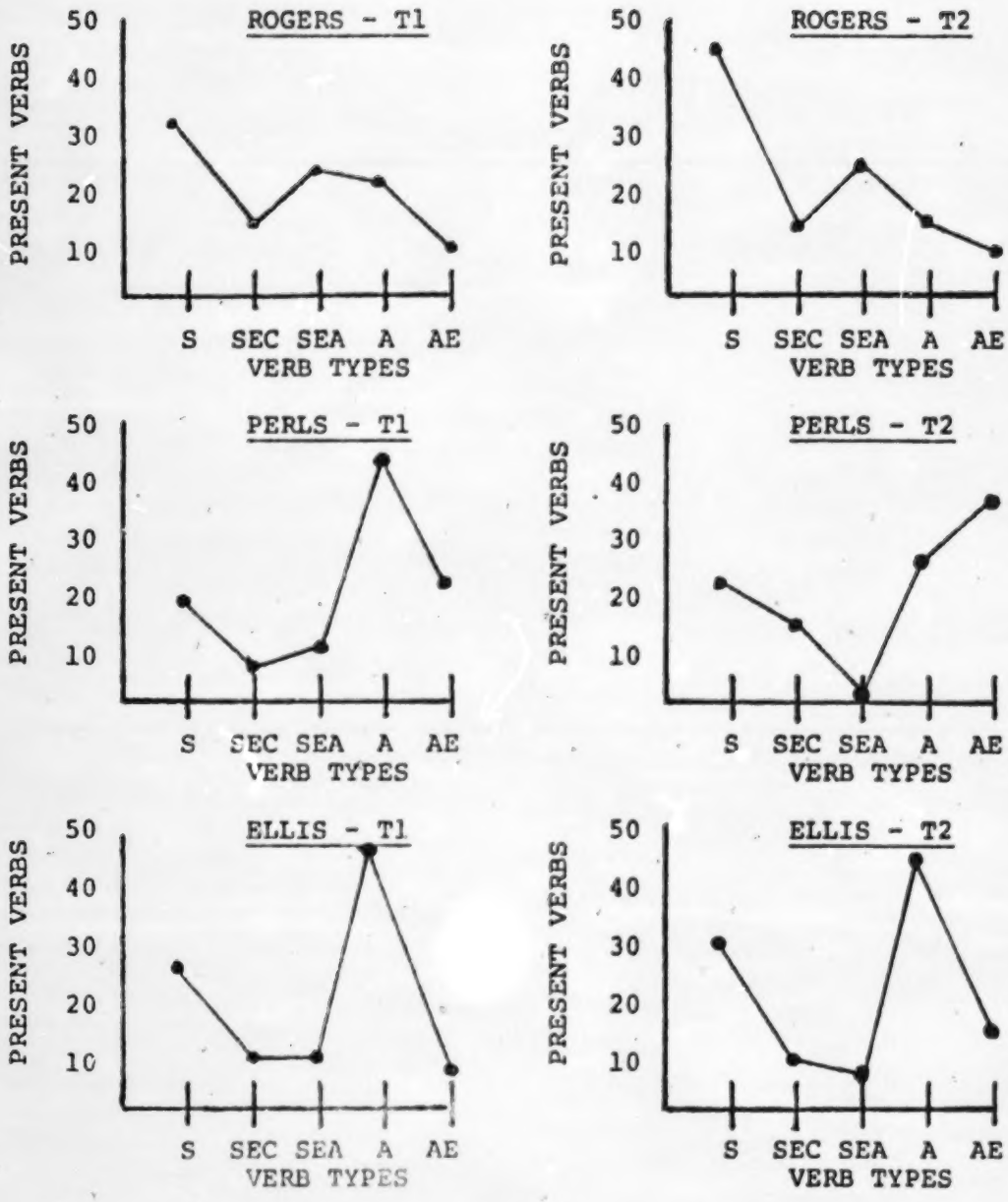

Figure 3. Percentage of verb use at two different time periods (T1 and T2) within the same interview.

(From Murray, 1978)